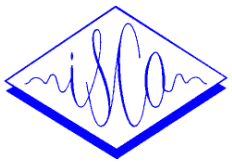

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

A REVERSE TURING TEST USING SPEECH

Greg Kochanski, Daniel Lopresti, and Chilin Shih

Bell Laboratories, Lucent Technologies
{gpk,dpl,cls}@research.bell-labs.com, 600 Mountain Ave, Murray Hill, NJ 07974

ABSTRACT

"Hackers" have written malicious programs to exploit online services intended for human users. As a result, service providers need a method to tell whether a web site is being accessed by a human or a machine. We expect a parallel scenario as spoken language interfaces become common.

In this paper, we describe a *Reverse Turing Test* (*i.e.*, an algorithm that can distinguish between humans and computers) using speech. We present a test that depends on the fact that human recognition of distorted speech is far more robust than automatic speech recognition techniques.

Our analysis of 18 different sets of distortions demonstrates that there are a variety of ways to make the problem hard for machines. In addition, humans and speech recognition systems make different kinds of mistakes, and this difference can be employed to improve discrimination.

1. INTRODUCTION

The use of the Internet as a means for distributing valuable information and content have also made it an attractive target for "hackers." Attacks involving malicious programs ("bots") that attempt to exploit online services intended for human users are already common. These programs consume resources, harass users, make attempts to guess passwords, steal and re-purpose copyrighted content, and invade privacy by reconstructing sensitive data from public views.

As a result, there is a need for automatic methods to tell whether the entity attempting to access a service is a human or a machine. This has come to be known as a *Reverse Turing Test*, or *RTT* (or sometimes a *Human Interactive Proof*). Unlike the test originally proposed by Alan Turing [1], a Reverse Turing Test is administered by a computer, not a human. For the test to be considered effective, humans should be able to pass it with ease, but machines should have a low probability of passing.

While it may seem that passwords and/or biometrics could provide a solution, one must keep in mind that these approaches require pre-registration. An RTT will work even if the user is anonymous and has never used the service before. The RTT problem is fundamentally different from validating a known user. Indeed, a Reverse Turing Test should be applied before authentication to prevent automated attacks on passwords.

Coates, Baird, and Fateman [2], and von Ahn *et al.* [3] have developed such a test based on a visual perception task. Their ideas are based on the observation that optical character recognition (OCR) systems are not as adept at reading degraded word images as humans are. That RTT is now used commercially to protect a free email service [4].

Speech-based services are proliferating because of their ease-of-use, portability, and potential for hands-free operation. Building a "bot" to navigate a spoken language interface is a tractable problem, especially if there is a fixed sequence of predefined prompts. Hence, we anticipate attacks on such systems and a similar need to prevent machines from abusing speech-based resources intended for human users. Previous work on speech RTTs can be found in [5, 6].

As in the vision RTTs, we exploit the fact that certain pattern recognition tasks are significantly harder for machines than they are for humans. We will use text-to-speech synthesis (TTS) to generate tests, and make use of the limitations of state-of-the art automatic speech recognition (ASR) technology. (We require only that RTTs cannot be broken cheaply or rapidly. Clearly, any RTT can be broken by hiring a human.)

In this paper, we present the core of a spoken language RTT. We assume a user with a cell phone; the test may consist of having the system speak: "Please enter the following digits on your keypad: ..." followed by a short, random digit string. The speech would be synthesized in a way that ASR is likely to fail the test, *e.g.*, by distorting the signal or adding "difficult" background noise to it after synthesis.

2. PROCEDURES

We designed a set of 18 RTTs based on different distortions of a speech signal (Table 2) to explore a broad range of possibilities. To test our RTTs, we synthesized 200 random 5-digit sequences using the Bell Labs English text-to-speech system [7], with the default male voice. We next distorted the signals and ran the Bell Labs speech recognition system [8] on them, with a grammar that allowed any digit se-

Name	Description	Error Ratio
white	• White Gaussian noise, 4000 Hz bandwidth.	15
buzz	• Sine waves at 700 Hz, 2100 Hz, 3500 Hz.	≥ 20
song	• Bell Labs Song (pop/rock).	> 15
chopin	• Chopin Polonaise for Piano No. 6, Op. 53.	> 20
chant	• Gregorian chant.	> 20
female	• Three overlapping instances of a female voice reading numbers.	> 20
pow	• 10 ms bursts of white Gaussian noise, repeated every 100 ms.	> 20
rnoise	• Every 100 ms, a section of the signal is replaced by white noise of the same RMS amplitude.	> 20
cell	• For each 30 ms window, decide if the data was lost. If so, and previous not lost, duplicate previous. If so, and previous is lost, set to zero. Simulates a bad cell phone channel.	> 10
echo	• Three echoes.	> 20
filter	• A random zero-phase filter.	> 5
distort	• Apply AGC on a 60 ms window, raise to a power, multiply by original amplitude.	≥ 20
mxa	• *rnoise* + *chopin*	> 20
mxb	• *song* + *echo*	> 20
mxc	• *white* + *pow*	≥ 20
mxd	• *female* + *buzz*	≥ 5
mxe	• *rnoise* + *distort*	≥ 20
mxf	• *filter* + *distort*	3

Table 1. *Tested RTTs. The error ratio is an estimate of the largest ratio of the ASR to human utterance error rates. The components of all mixtures were chosen to cause equal error rates for ASR utterance error rates near 85%.*

quence.

The recognition results were compared to the original digit sequences using approximate string matching ([9] and references therein) to identify added, missing, and substituted digits. From this we obtained, for each distortion, per-digit and per-utterance error rates, as well as a confusion matrix showing the frequency of each type of error.

A similar procedure was followed to test how well humans could recognize the signals. The authors each listened to 10 randomly-chosen digit sequences for each type of distortion. In these tests, the audio signals were presented twice with a one second pause in between. The subject then typed his/her interpretation and pressed return to listen to the next signal. The tests began at the most severe distortion of each type, and terminated when the subject correctly identified all 10 sequences. One set (*white*) received 67 repetitions per person for a more accurate confusion matrix.

The confusion matrices are scaled to make all diagonal confusion matrices identical. This step is necessary because the distance between raw confusion matrices is generally nonzero, even if there are no errors (*i.e.* the matrices are diagonal). This happens because we do not explicitly balance the frequencies of each digit, so that one test may see more instances of, say, "3" than another. Hence, we use a scaled confusion matrix, $S = P \cdot C \cdot Q$, where P and Q are diagonal matrices chosen so that the row- and column-sums of S are unity. While deletions and insertions are conventionally placed in the same matrix as substitutions they are actually a fundamentally different kind of error. Specifically, $C_{del,ins}$ is on the diagonal and is missing (zero). Thus, insertions and deletions must be treated differently in the scaling. We make an ad-hoc modification to C before scaling: we set $C_{del,ins} \leftarrow \sum_{i,j} C_{i,j}/N$ (where $N = 11$ is the number of rows and columns in C), and then set $S_{del,ins} \leftarrow 0$.

3. ANALYSIS: IS IT POSSIBLE TO DISTINGUISH HUMANS FROM MACHINES?

Human perception of speech in noisy environments is fairly robust. Normal-hearing listeners need a signal-to-noise ratio (SNR) of approximately 1.5 dB to recognize speech [10], while ASR systems require a much more favorable SNR of 5 to 15 dB [11]. Our test results show an even wider gap between human and ASR performance.

Figures 1-4 plot the error rates we measured in four experiments; these correspond to four distinct types of distortion: additive noise, deleting segments of the speech, adding echoes and filtering the signal. The four curves in each chart represent the error rates for ASR on a per-utterance (square) and per-symbol (diamond) basis along with per-utterance (X) and per-symbol (triangle) rates for humans.

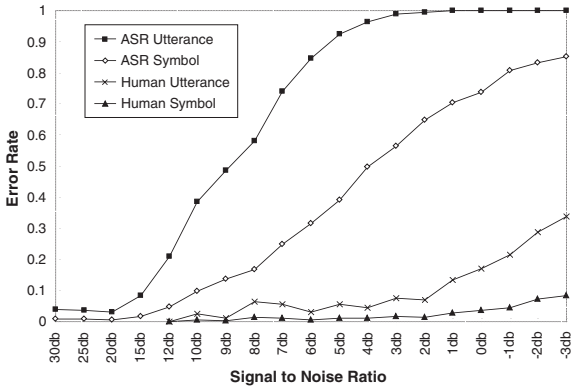

Fig. 1. Results for white noise experiment (*white*).

In figure 1, *white* is simply additive white noise. The x-axis shows SNR, and the y-axis represents the rate of recognition errors. In the ASR curves, each datum represents scores from 200 utterances or 1,000 digits, while in the human curves, each datum represents 201 utterances or 1,005 digits as pooled from the three listeners. ASR performance starts deteriorating when the SNR reaches 15 dB, and breaks down completely by 3 dB. Human performance is only starting to deteriorate at 0 dB SNR.

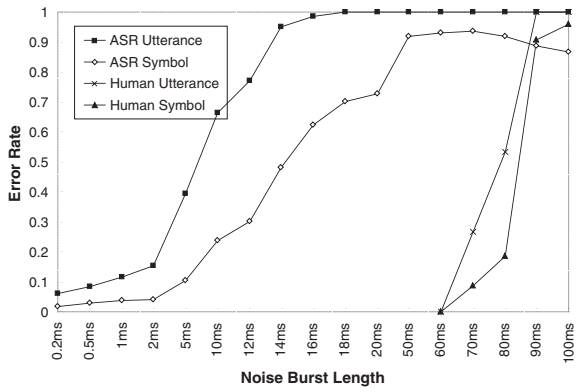

Fig. 2. Results for replacement noise experiment (*rnoise*).

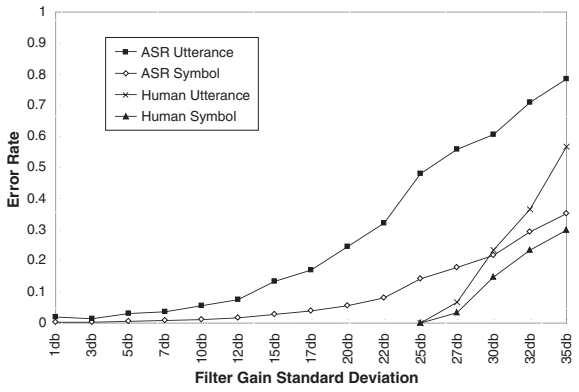

Fig. 4. Results for random filter experiment (*filter*).

Figure 2 shows the results for *rnoise*, which replaces a segment of speech with white noise every 100 ms, as we vary the length of the replaced segment. Human scores in this figure and in Figures 3-4 are based on 30 utterances (150 digits) per datum. As can be seen, ASR starts having problems when 2 ms (2%) of the speech is replaced. Amazingly, human recognition remains perfect at 60 ms (60% replacement), and 80% of the symbols are still correct even when 80% of the speech is missing.

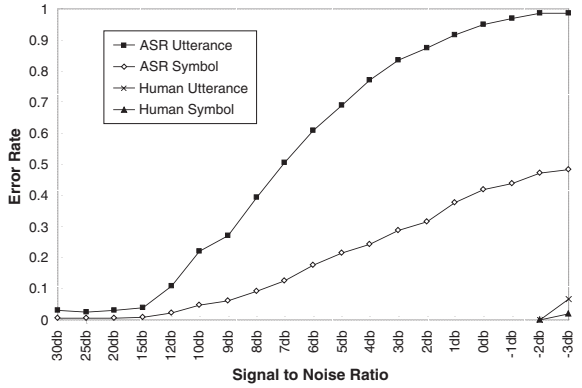

Fig. 3. Results for echo experiment (*echo*).

Our results for *echo*, which adds three echoes to the speech with delays of 60, 132, and 192 ms, are presented in Figure 3. The horizontal axis shows the relative amplitude of the first echo to the speech, while later echoes are 5 dB and 10 dB quieter. ASR performance starts declining when the SNR is 12 dB, while human performance is perfect until -3 dB. This behavior is typical of many of the other tests we performed.

Among our tests, *filter* (Figure 4) showed the smallest difference between ASR and humans. This is a zero-phase frequency domain filter, with a frequency response chosen randomly every 30 Hz. The control parameter sets the standard deviation of the gain, expressed in dB.

Our findings are that the gap between ASR and human performance appears to be wide enough to administer a robust Reverse Turing Test. In general, our tests show that humans can handle noise levels about 15 dB higher than ASR can. Likewise, humans can understand digit strings when more than half of the signal is missing, a point at which the machine already has a 100% error rate for utterances.

4. ANALYSIS: DO HUMANS MAKE THE SAME KINDS OF ERRORS AS MACHINES?

Not only is the error rate larger for an ASR system than for humans, but the pattern of errors is significantly different as well. Figure 5 shows a gray-scale plot of errors made by the machine and humans on the *white* data set. The figures represent noise levels chosen to give matching error rates.

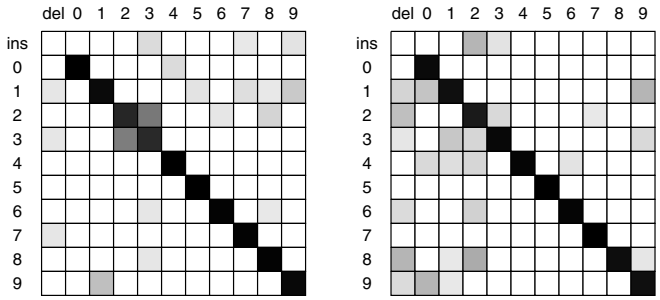

Fig. 5. Scaled confusion matrices at 35% error rate for humans (-3dB SNR, left) and machine (+10dB SNR, right).

Based on this observation, we could improve an RTT by looking at the kinds of errors that are made. For instance, humans readily confuse the digits “2” and “3” in the presence of white noise, so such an error should not be considered evidence that a machine is attempting to access the system. On the other hand, an “8” is never misunderstood as a “2” by humans but is sometimes by ASR, so such an error would suggest a machine's presence.

5. ANALYSIS: HOW MANY DISTINCT KINDS OF DISTORTION ARE THERE?

One obvious attack on this kind of RTT is to build a classifier that, working from the input signal, attempts to identify the distortion that was applied, and then sends the input to an ASR system trained specifically for that distortion. ASR can be tuned to work well in the presence of noise [12], but it first needs to be trained with a large representative corpus collected from the environment in question.

Assuming that it is possible to build the necessary classifier, the question then arises "How many different ASR systems would one need to train in order to break the RTT?" To make this question tractable, we approach it by examining the confusion matrices derived from the experiments we have performed; these will serve as a proxy for training and testing a large number of ASR systems.

To do this, we follow logic presented elsewhere in the context of OCR systems [13]. We assume that if the ASR confusion matrices corresponding to two distorted signals are different enough, then separate recognizers will be needed because the two signals are fundamentally different. To make a quantitative comparison, we need to define a distance measure: we use the 2-norm of the difference of the scaled confusion matrices, $D(\alpha,\beta)^2 = \sum_{i,j}(S_{i,j}(\alpha) - S_{i,j}(\beta))^2$, where α and β refer to two different distortions, and S is a function of how the speech is distorted.

We can calibrate our notion of "different enough" by picking an error rate at which the ASR system fails and measuring $D(\alpha, \mathrm{perfect})$ for various types of distortion at that rate. We choose noise levels that yield an utterance error rate of $85\% \pm 6\%$, with a symbol error rate of $34\% \pm 5\%$. The confusion matrices at these levels were found to be distance $D_{85} = 2.1 \pm 0.1$ from the perfect, error-free case. We assume that if the confusion matrices for any two distortions differ by this amount or more, an ASR system trained for one distortion will not be able to function on the other.

In our experiments, the average distance between a pair of distortions (excluding the mixtures) is 2.0 ± 0.9, quite close to D_{85}, so we would expect that, in general, an ASR system could not be trained to handle two different types of distortions simultaneously. The mixtures tend to be closer to their components, a distance of 1.8 ± 0.6 away, but this is still far enough to conclude that an ASR system trained on either component of the mixture would probably not perform well on the mixture itself. Consequently, it appears there are enough ways of generating fundamentally distinct distortions that the RTTs we have described would resist an attack using any single, well-trained ASR system.

6. CONCLUSIONS

In this paper, we have described our work towards building a speech-based Reverse Turing Test. We show that the gap between ASR and human performance is wide for a variety of noise effects, and that there are opportunities to exploit differences between the patterns of errors that humans and machines make. The RTT is a fundamentally new way of using speech synthesis and recognition technologies.

The authors thank Olivier Siohan for assistance.

7. REFERENCES

[1] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[2] A. L. Coates, H. S. Baird, and R. J. Fateman, "Pessimal print: A Reverse Turing Test," in *Proc. of the Sixth Intl. Conf. on Document Analysis and Recognition*, Seattle, WA, September 2001, pp. 1154–1158.

[3] L. von Ahn, M. Blum, J. Langford, and U. Manber, "The CAPTCHA project: Telling humans and computers apart (automatically)," October 2001, http://www.captcha.net/.

[4] Yahoo! Inc., ," March 2002, http://mail.yahoo.com.

[5] D. Lopresti, C. Shih, and G. Kochanski, "Human interactive proofs for spoken language interfaces," in *Proc. of the Workshop on Human Interactive Proofs*, Palo Alto, CA, January 2002, pp. 30–34.

[6] N. Chan, "Abstract of sound oriented CAPTCHA," in *Proc. of the Workshop on Human Interactive Proofs*, Palo Alto, CA, January 2002, p. 35.

[7] R. W. Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer, 1998.

[8] Q. Zhou and W. Chou, "An approach to continuous speech recognition based on layered self-adjusting decoding graph," in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997, pp. 1779–1782.

[9] D. Sankoff and B. Kruskal, Joseph, *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*, CSLI Publications, 2 edition, 1999, ISBN 1-57586-217-4. Originally Addison-Wesley, 1993.

[10] A. Stuart and D. P. Phillips, "Word recognition in continuous and interrupted broadband noise by young normal-hearing, older normal-hearing, and presbyacusic listeners," *Ear & Hearing*, vol. 17, pp. 478–489, 1996.

[11] E. Woudenberg, F. K. Soong, and J. E. West, "Acoustic echo cancellation for hands-free ASR applications in noise," in *Proc. of the Workshop on Acoustic Echo and Noise Control*, 1999, pp. 160–163.

[12] P. J. Moreno, B. Raj, and R. M. Stern, "Data-driven environmental compensation for speech recognition: A unified approach," *Speech Communication*, vol. 24, pp. 267–285, 1998.

[13] Y. Li, D. Lopresti, G. Nagy, and A. Tomkins, "Validation of image defect models for optical character recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 99–108, 1996.